# A New Graphical Representation of DNA Sequences Using Symmetrical Vector Assignment

Kyohei Yamaguchi, Satoshi Mizuta*

Graduate School of Science and Technology, Hirosaki University3 Bunkyo-cho, Hirosaki, Aomori 036-8561, Japan

*slmizu@cc.hirosaki-u.ac.jp

Received Feb 26, 2014; Accepted Mar 11, 2014; Published Jun 9, 2014


***Abstract***

Analyzing the similarities between genomic sequences is one of the principal methods used to investigate the evolutionary relationships between species. For relatively short sequences, such as nucleotide sequences of genes or amino acid sequences of proteins, alignment is widely used to evaluate the sequence similarity. However, the alignment is not practical for comparing very long sequences, such as genome sequences, due to its time-consuming nature. In this article, we propose a new method for graphical representation of DNA sequences, which falls into one of the major categories of alignment-free sequence comparison. We introduce a practical method for the numerical conversion of DNA sequences, in which we assign three-dimensional vectors in a symmetrical manner to the bases of genome sequences. We confirm the usefulness of our method in terms of the intuitive assessment of sequence similarities.

***Keywords***

*Alignment-free; Sequence Comparison; Mitochondrial Genome; 3D Graph*

## Introduction

In comparative genomics, comparing genome sequences is one of the main tasks because sequence similarities strongly reflect the evolutionary relationships between the corresponding species. In addition, with the introduction of next-generation sequencing technologies, the demand for rapid comparisons of massive amounts of long sequences has increased in recent years.

Sequence alignment (Smith and Waterman 1981; Needleman and Wunsch 1970) is generally used to compare relatively short sequences, such as nucleotide sequences of genes or amino acid sequences of proteins. The time complexity of the sequence alignment is $O(N^2)$ for sequences of length $N$, which indicates that the sequence alignment is very time-consuming when $N$ is extremely large (i.e. for instance, alignment of whole genome sequences). Therefore, along with improvements in alignment-based methods, alignment-free methods are actively studied by many researchers to perform comparison between such long sequences.

Graphical representation of DNA sequences is one of the alignment-free methods, which provide visual inspection of DNA sequences and makeit possible to compare DNA sequences instantly. Various schemes for the graphical representation have been proposed by several authors based on the projection of DNA sequences on 2D (Qi, Li, and Qi 2011; Huang et al. 2011; Yu et al. 2010; Zhang 2009; Randić 2008; Qi and Qi 2007; Bielińska-Wąz et al. 2007a; Bielińska-Wąz et al. 2007b; Zhang and Chen 2006; Liu et al. 2006; Song and Tang 2005; Liao, Tan, and Ding 2005; Liao and Wang 2004d; Randić et al. 2003a; Randić et al. 2003b; Wu et al. 2003; Liu et al. 2002; Randić and Vračko 2000; Nandy 1994; Jeffrey 1990; Gates 1985), 3D (Xie and Mo 2011; Yu and Sun 2010; Yu, Sun, and Wang 2009; Cao, Liao, and Li 2008; Qi, Wen, and Qi 2007; Qi and Fan 2007; Liao and Ding 2006; Yao, Nan, and Wang 2005; Liao and Wang 2004a; Liao and Wang 2004b; Balaban, Plavšić, and Randić 2003; Zhang, Zhang, and Ou 2003; Randić et al. 2000; Hamori 1985; Hamori and Ruskin 1983), or higher dimensional spaces (Liao et al. 2007; Chi and Ding 2005; Liao and Wang 2004; Randić and Balaban 2003). The basic procedure is common in almost all of the above-mentioned schemes: numerical conversion of bases of DNA sequences, consecutive mapping of the converted bases on a certain dimensional space to draw a graph, and estimation of

the similarity between the graphs. In this study, we propose a novel method for graphical representation of DNA sequences on a 3D space, in which we adopt symmetrical vector assignments, and introduce weighting in numerical conversion.

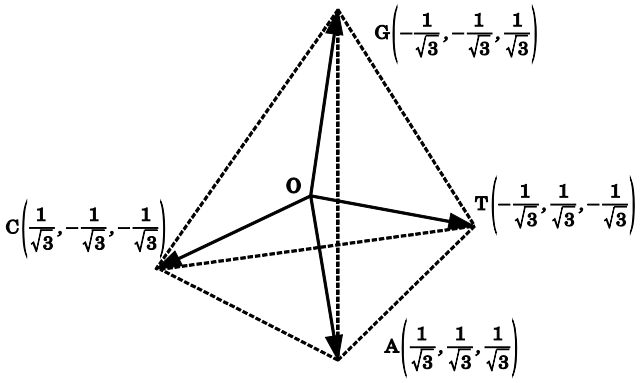


FIGURE 1 VECTORS ASSIGNED TO FOUR TYPES OF BASES.

## Method

### *Symmetrical Vector Assignment*

We assign distinct vectors of a certain dimension to each of four types of bases, A, T, G, and C, for numerical conversion. By connecting the vectors corresponding to the bases extracted one by one from the head of the target genome sequence, we can perform its graphical representation. If we map genome sequences on a 2D space, the interrelationship between the resultant graphs may change according to the arrangement of the vectors due to their asymmetrical nature (i.e. not all the distances between the end points of each pair of vectors out of four can be equal). In this study, therefore, we map genome sequences on a 3D space using vectors represented by the vertices of a regular tetrahedron with edges of length 1 (FIG. 1). Here, we should emphasize that all the arrangements of the four bases on the vertices can be mutually transformed by rotation and/or space inversion not affecting the distances between the resultant graphs because the distance we will define is invariant under the rotation and the space inversion (see *Distance measure between sequences*); therefore, only the configuration shown in FIG. 1 needs to be considered.

### *Weighting Factors*

In order to evaluate effectively the information that each base in a genome sequence conveys, we assigned weighting factors to the vectors according to the appearance probabilities of the corresponding bases. We used self-information of the appearance of each kind of base for the weighting factor. Let $P$ be the probability that a certain event occurs, thus the self-information $I$ for the occurrence of the event is expressed by

$$I = -\log P. \tag{1}$$

Here, we take the conditional probability of the occurrence of each base as $P$. A conditional probability is the probability that an event occurs given that another event has already occurred. For example, the conditional probability $P(\mathrm{A}|\mathrm{GC})$ measures the probability that base A appears after a pair of bases GC, which is computed by

$$P(\mathrm{A}\,|\,\mathrm{GC}) = \frac{\#\mathrm{GCA}}{\#\mathrm{GCA}+\#\mathrm{GCT}+\#\mathrm{GCG}+\#\mathrm{GCC}}, \tag{2}$$

where #W (W= "GCA", "GCT", . . . ) represents the number of occurrences of string W.

As for the string length for calculating the conditional probability, we paid attention to the fact that amino acids are encoded by codons (i.e. triplets of bases) in genome sequences, and we used length three to get some information of the coding regions of the genome sequences, although the differences were not so large among the results obtained with different lengths (data not shown).

We computed the conditional probabilities using all the genome sequences analyzed in this study. TABLE 1 shows the weighting factors computed according to the above mentioned procedures based on tri-nucleotides.

TABLE 1 WEIGHTING FACTORS FOR BASES

| Preceding sequence | Base | | | |
|---|---|---|---|---|
| | A | G | C | T |
| AA | 1.13 | 2.05 | 1.32 | 1.27 |
| AG | 1.44 | 1.53 | 1.10 | 1.54 |
| AC | 1.15 | 2.32 | 1.25 | 1.21 |
| AT | 1.17 | 2.00 | 1.35 | 1.22 |
| GA | 1.08 | 1.61 | 1.45 | 1.49 |
| GG | 1.08 | 1.72 | 1.31 | 1.55 |
| GC | 1.18 | 2.76 | 1.04 | 1.29 |
| GT | 0.93 | 1.91 | 1.54 | 1.41 |
| CA | 1.16 | 1.99 | 1.34 | 1.25 |
| CG | 1.25 | 1.67 | 1.28 | 1.40 |
| CC | 1.24 | 2.51 | 1.19 | 1.13 |
| CT | 0.95 | 2.11 | 1.42 | 1.39 |
| TA | 1.18 | 1.77 | 1.42 | 1.27 |
| TG | 0.99 | 1.61 | 1.48 | 1.61 |
| TC | 1.07 | 2.36 | 1.26 | 1.28 |
| TT | 1.06 | 2.10 | 1.32 | 1.33 |

### *Graphical Representation*

Graphical representation of a genome sequence is

performed by connecting sequentially the weighted vectors corresponding to the bases in the genome sequence. In drawing a graph, the start point is set to the origin of the 3D space. Here, we show you a simple example of the procedure. Let "GATCA" be a nucleotide sequence. We begin the graphical representation with the third base 'T' because we need two preceding bases to assign the weighting factor. The corresponding vector to 'T' is $(-1/\sqrt{3}, 1/\sqrt{3}, -1/\sqrt{3})$ (FIG. 1) and the weighting factor for 'T' of "GAT" is 1.49 (TABLE 1). Then the coordinate value of 'T' is calculated to be 1.49 $(-1/\sqrt{3}, 1/\sqrt{3}, -1/\sqrt{3}) = (-0.86, 0.86, -0.86)$. Similarly, the weighted vector for the next base 'C' is calculated to be 1.35 $(1/\sqrt{3}, -1/\sqrt{3}, -1/\sqrt{3}) = (0.78, -0.78, -0.78)$ and is added up to the above coordinate value; that is, $(-0.08, 0.08, -1.64)$. This procedure is continued to the end of the sequence. Thus, the graphical representation of "GATCA" is completed (FIG. 2).

When the appearances of the resultant graphs are similar for some genome sequences, the corresponding species can be considered to be closely related to each other, and when completely dissimilar, the corresponding species can be considered to be distantly related to each other. Note again that, due to the symmetric properties of the vector assignment (see Symmetrical vector assignment), the basic features of the resultant graphs are independent of the arrangement of the vectors, although the appearances of the graphs may change according to the arrangement.

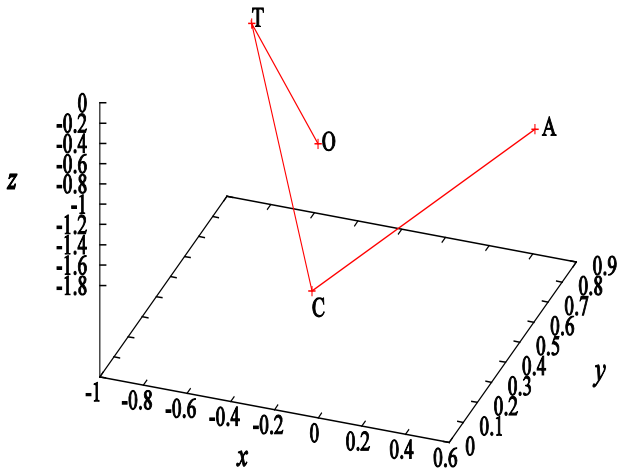


FIGURE 2 GRAPHICAL REPRESENTATION OF "GATCA"

### *Distance Measure Between Sequences*

We need to define the distance between the resultant graphs to evaluate quantitatively the similarities between the corresponding sequences. We divided each sequence into four segments of equal length, and created a 12-dimensional feature vector from the coordinate values of the four points—the three boundary points of the segments and the terminal. We then defined the distance between the sequences using the Euclidean distance between the feature vectors. That is, the square of distance L between two sequences is calculated by the following formula:

$$L^2 = \sum_{i=1}^{4}(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2, \quad (3)$$

where $x_i(x_i')$, $y_i(y_i')$, and $z_i(z_i')$ are the coordinate values of the $i$-th sampling point of the sequence; $i$=1,2,3 corresponds to the three boundary points of the four segments of the divided sequence, and $i$=4 corresponds to the terminal of the sequence.

We attempted two other variations in the number of sampling points for calculating the distances: all the points of the sequence and the terminal point only. As a result, we obtained the best performance when using the four-point sampling. Therefore, we consider only the four-point sampling in calculating the distances hereafter.

## Results And Discussion

### *Data Set*

The nucleotide sequences of mitochondrial genomesfor 38 mammals were downloaded from GenBank and used for analysis (TABLE 2).

### *Graphs And Effects Of Weighting*

Initially, we compared the graphs of closely related species—common chimpanzee and pygmy chimpanzee—in FIG. 3 (upper panel). We can find that the appearances of the graphs are very similar. The lower panel of FIG. 3 shows the same graphs but without weighting in numerical conversion of the sequences. It is evident in this figure that the weighting emphasizes the characteristics of the graphs, and makes it possible to distinguish between the graphs of close relatives. Next, we compared the graphs of distant relatives—dog and common chimpanzee—in FIG. 4. We can find that the appearances of the graphs are quite different. These observations support the usefulness of our new method of graphical representation in terms of intuitive assessment of sequenc similarities.

### *Phylogenetic Tree*

We calculated the distances between all pairs of

species listed in TABLE 2 using Eq.(3), and constructed a distance matrix. FIG. 5 shows the phylogenetic tree created from the distance matrix. The tree was drawn by the statistical analysis software R based on the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method.

TABEL 2 LIST OF MITOCHONDRIAL GENOMES OF 38 MAMMALS ANALYZED

| Species | Common name | Accession No. |
|---|---|---|
| *Homo sapiens* | Human | V00662 |
| *Pan paniscus* | Common chimpanzee | D38113 |
| *Pan troglodytes* | Pygmy chimpanzee | D38116 |
| *Gorilla gorilla* | Gorilla | D38114 |
| *Pongo pugmaeus* | Orangutan | D38115 |
| *Hylobates lar* | Gibbon | X99256 |
| *Papio hamadryas* | Baboon | Y18001 |
| *Equus caballus* | Horse | X79547 |
| *Ceratotherium simum* | White rhinoceros | Y07726 |
| *Rhinoceros unicornis* | India rhinoceros | X97336 |
| *Phoca vitulina* | Harbor seal | X63726 |
| *Halichoerus grypus* | Gray seal | X72004 |
| *Felis catus* | Cat | U20753 |
| *Panthera tigris* | Tiger | EF551003 |
| *Panthera pardus* | Leopard | EF551002 |
| *Balenoptera physalus* | Fin whale | X61145 |
| *Balenoptera musculus* | Blue whale | X72204 |
| *Bos taurus* | Cow | V00654 |
| *Bubalus bubalis* | Buffalo | AY488491 |
| *Rattus norvegicus* | Norway rat | X14848 |
| *Mus musculus* | Mouse | V00711 |
| *Dudelphis virginiana* | Opossum | Z29753 |
| *Macropus robustus* | Wallaroo | Y10524 |
| *Ornithorhyncus anatinus* | Platypus | X83427 |
| *Canis lupus familiaris* | Dog | U96639 |
| *Canis lupus chanco* | Wolf | EU442884 |
| *Sus scrofa* | Pig | AJ002189 |
| *Ovis aries* | Sheep | AF010406 |
| *Loxodonta africana* | African elephant | AJ224821 |
| *Elephas maximus* | Asiatic elephant | DQ316068 |
| *Ursus thibetanus mupinensis* | Black bear | DQ402478 |
| *Ursus arctos* | Brown bear | AF303110 |
| *Ursus maritimus* | Polar bear | AF303111 |
| *Oryctolagus cuniculus* | Rabbit | AJ001588 |
| *Erinaceus europaeus* | Hedgehog | X88898 |
| *Microtus kikuchii* | Vole | AF348082 |
| *Sciurus vulgaris* | Squirrel | AJ238588 |

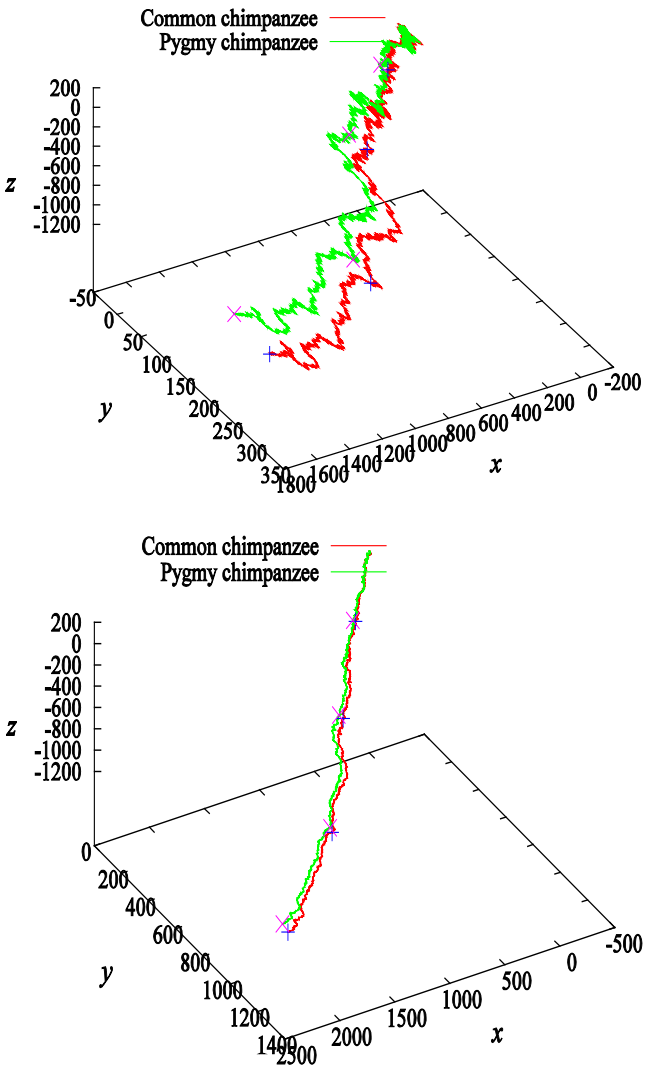


FIGURE 3 GRAPHS OF CLOSE RELATIVES WITH (UPPER PANEL) AND WITHOUT WEIGHTING (LOWER PANEL). SYMBOLS '×' AND '+' SHOW THE SAMPLING POINTS TO CALCULATE THE DISTANCE BETWEEN THE SEQUENCES

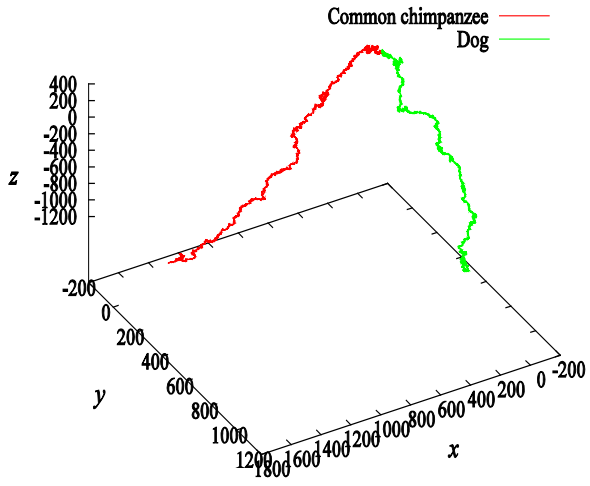


FIGURE 4 GRAPHS OF DISTANT RELATIVES—DOG AND COMMON CHIMPANZEE

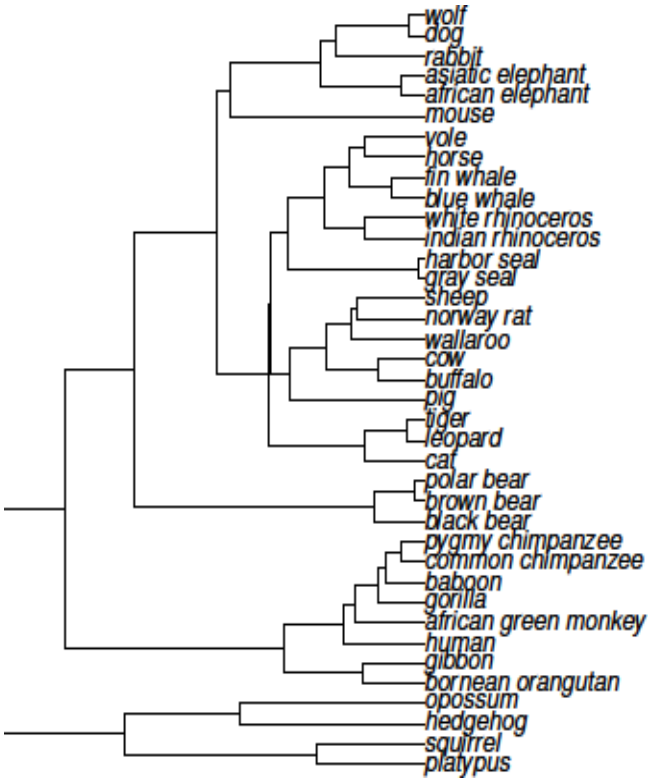


FIGURE 5 PHYLOGENETIC TREE BASED ON THE UPGMA METHOD REPRESENTING ALL THE SPECIES ANALYZED

The configuration of the phylogenetic tree is largely inagreement with those in (Huang et al. 2011) and (Yuet al. 2010), with primates, bears, elephants, seals, and cats (cat, tiger, and leopard) being located in their respective clusters. However, certain species seem to be located on inappropriate positions. For example, the three rodent species (Norway rat, mouse, and vole) are separated from each other. One of the major causes of this anomaly seems to be the way of definition of the distance between sequences (see *Distance measure between sequences*). In our method, we take four sampling points of each genome sequence for calculating the distances. However, the configuration of these sampling points depends on the start point of the sequence. Currently, we take the head of the sequence data as the start point, although the start point of a mitochondrial genome is not apparently determined due to its circular form. We are now engaged in improving the definition of the distance between sequences considering the start point.

FIG. 6 shows the graphs of several clusters of species, which are closely located in the phylogenetic tree (FIG. 5). The appearances of the graphs in each cluster look similar, whereas those of the graphs between different clusters are highly dissimilar. This observation confirms that the phylogenetic tree properly reflects the similarities between the graphs.

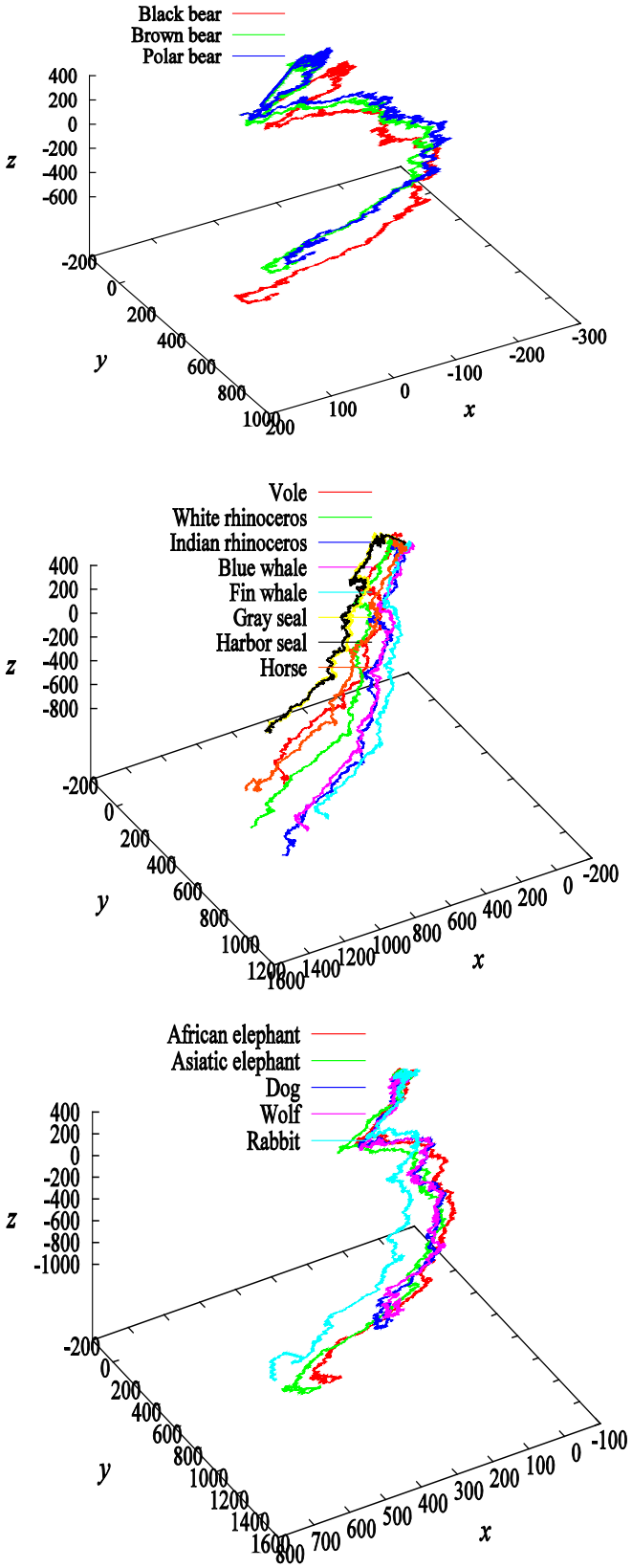


FIGURE 6 GRAPHS OF THE SPECIES CLOSELY LOCATED IN THE PHYLOGENETIC TREE

## Conclusion

We proposed a novel method for graphical

representation of DNA sequences. In this method, we assigned three-dimensional vectors represented by the vertices of a regular tetrahedron to each base, and gave weighting to the vectors based on the self-information of the appearance of the corresponding bases. Our method has a significant feature in that the quantitative outcomes with respect to sequence similarities are independent of the arrangement of the vectors due to its symmetric nature.

By comparing the graphs of close and distant relatives, we confirmed the effects of weighting and the usefulness of our method in terms of the intuitive assessment of sequence similarities. Furthermore, we defined the distance between graphs to evaluate sequence similarities quantitatively, and constructed a distance matrix including all the species analyzed to create the phylogenetic tree based on the distance matrix with the UPGMA method. We classified the species into some clusters by gathering the species closely located in the phylogenetic tree to each other, and compared the appearances of the corresponding graphs within and between the clusters. The appearances of the graphs of the species in each cluster are similar to each other, whereas those between different clusters are dissimilar. We therefore conclude that our method is effective for evaluating sequence similarities on an intuitive basis. However, our distance measure requires some refinements since certain species were located at the incorrect positions in the phylogenetic tree. We are now improving the definition of the distance between sequences in terms of identifying the appropriate start point of mitochondrial genome sequences.

## REFERENCES

Balaban, Alexandru T., Dejan Plavšić, and Milan Randić. 2003. "DNA invariants based on nonoverlapping triplets of nucleotide bases." Chemical Physics Letters 379 (1-2): 147–154.

Bielińska-Wąz, Dorota, Timothy Clark, Piotr Wąz, Wiesław Nowak, and Ashesh Nandy. 2007a. "2Ddynamic representation of DNA sequences." Chemical Physics Letters 442 (1-3): 140–144.

Bielińska-Wąz, Dorota, Wiesław Nowak, Piotr Wąz, Ashesh Nandy, and Timothy Clark. 2007b. "Distribution moments of 2D-graphs as descriptors of DNA sequences." Chemical Physics Letters 443 (4-6): 408–413.

Cao, Zhi, Bo Liao, and Renfa Li. 2008. "A group of 3D graphical representation of DNA sequences based on dual nucleotides." International Journal of Quantum Chemistry 108 (9): 1485–1490.

Chi, Rui, and Kequan Ding. 2005. "Novel 4D numerical representation of DNA sequences." Chemical Physics Letters 407 (1-3): 63–67.

Gates, M. A. 1985. "Simpler DNA sequence representations." Nature 316:219–219.

Hamori, E, and J Ruskin. 1983. "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences." Journal of Biological Chemistry 258 (2): 1318–27.

Hamori, Eugene. 1985. "Novel DNA sequence representations." Nature 314:585–585.

Huang, Guohua, Houqing Zhou, Yongfan Li, and Lixin Xu. 2011. "Alignment-free comparison of genome sequences by a new numerical characterization." Journal of Theoretical Biology 281 (1): 107–112.

Jeffrey, H. Joel. 1990. "Chaos game representation of gene structure." Nucleic Acids Research 18 (8): 2163–2170.

Liao, Bo, and Kequan Ding. 2006. "A 3D graphical representation of DNA sequences and its application." Theoretical Computer Science 358 (1): 56–64.

Liao, Bo, and Tian-Ming Wang. 2004a. "3-D graphical representation of DNA sequences and their numerical characterization." Journal of Molecular Structure: THEOCHEM 681 (1-3): 209–212.

Liao, Bo, and Tian-Ming Wang. 2004b. "Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation." Chemical Physics Letters 388 (1-3): 195–200.

Liao, Bo, and Tian-Ming Wang. 2004c. "Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases." Journal of Chemical Information and Computer Sciences 44 (5): 1666–1670.

Liao, Bo, and Tian-Ming Wang. 2004d. "New 2D graphical representation of DNA sequences." Journal of Computational Chemistry 25 (11): 1364–1368.

Liao, Bo, Mingshu Tan, and Kequan Ding. 2005. "Application of 2-D graphical representation of DNA sequence." Chemical Physics Letters 414 (4-6): 296–300.

Liao, Bo, Renfa Li, Wen Zhu, and Xuyu Xiang. 2007. "On the

Similarity of DNA Primary Sequences Based on 5-D Representation." Journal of Mathematical Chemistry 42 (1): 47–57.

Liu, Xiao Qing, Qi Dai, Zhilong Xiu, and Tianming Wang. 2006. "PNN-curve: A new 2D graphical representation of DNA sequences and its application." Journal of Theoretical Biology 243 (4): 555–561.

Liu, Yachun, Xiaofeng Guo, Jin Xu, Linqiang Pan, and Shiying Wang. 2002. "Some Notes on 2-D Graphical Representation of DNA Sequence." Journal of Chemical Information and Computer Sciences 42 (3): 529–533.

Nandy, A. 1994. "A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes." Current Science 66:309–314.

Needleman, Saul B., and Christian D. Wunsch. 1970."A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." J. Mol. Biol. 48:443–453.

Qi, Xiao-Qin, Jie Wen, and Zhao-Hui Qi. 2007. "New 3D graphical representation of DNA sequence based on dual nucleotides." Journal of Theoretical Biology 249 (4): 681–690.

Qi, Zhao-Hui, and Tong-Rang Fan. 2007. "PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization." Chemical Physics Letters 442 (4-6): 434–440.

Qi, Zhao-Hui, and Xiao-Qin Qi. 2007. "Novel 2D graphical representation of DNA sequence based on dual nucleotides." Chemical Physics Letters 440 (1-3): 139–144.

Qi, Zhao-Hui, Ling Li, and Xiao-Qin Qi. 2011. "Using Huffman coding method to visualize and analyse DNA sequences." Journal of Computational Chemistry 32 (15): 3233–3240.

Randić, M., M. Vračko, A. Nandy, and S. C. Basak. 2000. "On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization." Journal of Chemical Information and Computer Sciences 40 (5): 1235–1244.

Randić, Milan, and Alexandru T. Balaban. 2003. "On A Four-Dimensional Representation of DNA Primary Sequences." Journal of Chemical Information and Computer Sciences 43 (2): 532–539.

Randić, Milan, and Marjan Vračko. 2000. "On the Similarity of DNA Primary Sequences." Journal of Chemical Information and Computer Sciences 40 (3): 599–606.

Randić, Milan, Marjan Vračko, Nella Lerš, and Dejan Plavšić. 2003a. "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation." Chemical Physics Letters 371 (1-2): 202–207.

Randić, Milan, Marjan Vračko, Nella Lerš, and Dejan Plavšić. 2003b. "Novel 2-D graphical representation of DNA sequences and their numerical characterization." Chemical Physics Letters 368 (1-2): 1–6.

Randić, Milan. 2008. "Another look at the chaos-game representation of DNA." Chemical Physics Letters 456 (1-3): 84–88.

Smith, T. F., and M. S. Waterman. 1981. "Identification of common molecular subsequences." Journal of Molecular Biology 147:195–197.

Song, Jie, and Huanwen Tang. 2005. "A new 2-D graphical representation of DNA sequences and their numerical characterization." Journal of Biochemical and Biophysical Methods 63 (3): 228–239.

Wu, Yonghui, Alan Wee-Chung Liew, Hong Yan, and Mengsu Yang. 2003. "DB-Curve: a novel 2D method of DNA sequence visualization and representation." Chemical Physics Letters 367 (1-2): 170–176.

Xie, Guosen, and Zhongxi Mo. 2011. "Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications." Journal of Theoretical Biology 269 (1): 123–130.

Yao, Yu-Hua, Xu-Ying Nan, and Tian-Ming Wang. 2005. "Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation." Chemical Physics Letters 411 (1-3): 248–255.

Yu, Chenglong, Qian Liang, Changchuan Yin, Rong L. He, and Stephen S.-T. Yau. 2010. "A Novel Construction of Genome Space with Biological Geometry." DNA Research 17 (3): 155–168.

Yu, Jia-Feng, and Xiao Sun. 2010. "Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence." Journal of Computational Chemistry 31 (11): 2126–2135.

Yu, Jia-Feng, Xiao Sun, and Ji-Hua Wang. 2009. "TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications." Journal of

Theoretical Biology 261 (3): 459–468.

Zhang, Chun-Ting, Ren Zhang, and Hong-Yu Ou. 2003. "The Z curve database: a graphic representation of genome sequences." Bioinformatics 19 (5): 593–599.

Zhang, Yusen, and Wei Chen. 2006. "Invariants of DNA sequences based on 2DD-curves." Journal of Theoretical Biology 242 (2): 382–388.

Zhang, Zhu-Jin. 2009. "DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences." Bioinformatics 25 (9): 1112–1117.